# Robustness and User-Perceived Value of Popularity Calibration in Music Recommendation: A User Study

OLEG LESOTA, Johannes Kepler University Linz, Austria
GUSTAVO ESCOBEDO, Johannes Kepler University Linz, Austria
BRUCE FERWERDA, Jönköping University, Sweden
SIMONE KOPEINIK, Know-Center GmbH, Austria
DOMINIK KOWALD, Know Center Research GmbH and University of Graz, Austria
ELISABETH LEX, Graz University of Technology, Austria
MARKUS SCHEDL, Johannes Kepler University Linz and Linz Institute of Technology, Austria

Popularity calibration in recommender systems has been studied both as a form of user-centered personalization and as an indicator of popularity bias. Most existing work evaluates calibration through offline metrics, often assuming that users prefer recommendation lists whose popularity distribution matches their historical consumption profile. However, user studies on calibration remain limited, and existing findings suggest that calibrated recommendations do not necessarily have a strong effect on user experience. Moreover, although prior work has shown that calibration metrics can correlate with users' perceptions of recommendation lists, the robustness of this relation remains unclear under different levels of item familiarity and incomplete user-history information. In this work, we study the perceived value and measurement reliability of popularity calibration in music recommendation. We construct personalized track lists from users' recent listening histories and use a controlled naive recommender to create lists with different popularity compositions: `HighPop-Heavy`, `LowPop-Heavy`, and `Calibrated`. We investigate whether users perceive differences between these lists, whether calibrated lists are preferred, how robust JSD-based popularity calibration is under different familiarity and history-availability conditions, and how computational popularity labels align with users' own popularity judgments. Our results show that users perceive differences in popularity composition, but do not clearly prefer calibrated lists. We further find that the relation between JSD and perceived popularity depends on item familiarity, list composition, and available user history, while computational and user-judged popularity labels only weakly align. These findings contribute to a more critical understanding of popularity calibration as both an offline metric and a user-facing construct.

Additional Key Words and Phrases: recommender systems, music recommendation, user study, popularity bias, popularity calibration, miscalibration, metrics, ecological validity

**ACM Reference Format:**
Oleg Lesota, Gustavo Escobedo, Bruce Ferwerda, Simone Kopeinik, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2026. Robustness and User-Perceived Value of Popularity Calibration in Music Recommendation: A User Study. *ACM Trans. Recomm. Syst.* ?, ?, Article ? (December 2026), 18 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' addresses: Oleg Lesota, oleg.lesota@jku.at, Johannes Kepler University Linz, Linz, Austria; Gustavo Escobedo, gustavo.escobedo@jku.at, Johannes Kepler University Linz, Linz, Austria; Bruce Ferwerda, bruce.ferwerda@ju.se, Jönköping University, Department of Computer Science and Informatics, Jönköping, Sweden; Simone Kopeinik, skopeinik@know-center.at, Know-Center GmbH, Graz, Austria; Dominik Kowald, dkowald@know-center.at, Know Center Research GmbH and University of Graz, Graz, Austria; Elisabeth Lex, elisabeth.lex@tugraz.at, Graz University of Technology, Graz, Austria; Markus Schedl, markus.schedl@jku.at, Johannes Kepler University Linz and Linz Institute of Technology , Linz, Austria.


2770-6699/2026/12-ART?
https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Popularity calibration has been proposed as a user-centered way to study popularity bias in recommender systems [2, 5]. Instead of only measuring whether a system over-recommends popular items overall, calibration compares the popularity distribution of a user's past interactions with that of the recommendation list. This distinction is important because the same list can be calibrated for one user and miscalibrated for another: a list dominated by popular tracks may match the history of a mainstream-oriented listener, while it may be miscalibrated for a listener whose history contains more niche or long-tail music [2, 15, 20].

This framing makes popularity calibration attractive as a personalization and bias-mitigation concept, but it also introduces an important assumption: the user's observed listening history is treated as a meaningful target distribution for future recommendations. In practice, this target distribution depends on several design choices, including how item popularity is computed, how popularity bins are defined, and which parts of the user's history are available for calibration [2, 15, 20]. These choices are not merely technical, because they affect what it means for a recommendation list to be considered calibrated and what kind of user preference the metric is assumed to represent.

A further challenge is item familiarity. Prior work on perceived popularity bias in music recommendation suggests that users may not always recognize computational differences in popularity distributions, and that perceived popularity can be shaped by artist- or track-level familiarity [11]. This matters for user studies on calibration: if users evaluate unfamiliar tracks, their judgments may reflect lack of recognition rather than the popularity distribution of the list. In the present study, we therefore reduce the role of unfamiliarity by constructing recommendation lists from tracks users have recently consumed. This allows us to study perceived differences between popularity distributions under conditions where users are more likely to recognize the presented items.

Popularity calibration also implicitly treats a user's listening history as evidence of a popularity-related preference [2, 5]. However, in music recommendation, the relevant history is not obvious. A user's complete listening history may span many years, while recent listening may better reflect current taste. Moreover, not all tracks in a user's history may be available in the catalog used for computing popularity, and popularity estimates may differ depending on the dataset and time period considered [15, 20]. This raises the question of how robust popularity calibration is when the target distribution is estimated from incomplete or differently sized portions of user history.

Taken together, these issues underline the importance of studying popularity calibration not only as an offline metric, but also as a user-facing construct. Prior work has already started to compare computational and perceived popularity bias or miscalibration in music recommendation [11, 19]. We build on this line of work by examining the effects of item familiarity and incomplete user history. If calibration is intended to represent a user-centered notion of recommendation quality, then users should perceive meaningful differences between differently calibrated lists. Furthermore, calibrated lists should provide measurable value to users, and the underlying metric should remain robust in the face of realistic limitations in user-history data.

In this work, we conduct a user study in which participants evaluate music track lists with different popularity compositions. Rather than relying on standard recommender algorithms, we use a controlled naive recommender that constructs lists from tracks in users' recent listening histories. This design allows us to compare lists dominated by high-popularity tracks, lists dominated by low-popularity tracks, and popularity-calibrated lists while reducing the confounding effect of unfamiliar items. We then relate participants' track-level and list-level judgments to offline popularity-calibration measurements, including Jensen–Shannon divergence (JSD) between the popularity distribution of the list and the user's estimated historical popularity profile. More precisely, we address the following research questions:

- **RQ1:** Do users perceive differences in popularity levels between recommendation lists generated with different calibration targets?
- **RQ2:** Do users prefer popularity-calibrated recommendation lists over lists dominated by high- or low-popularity items?
- **RQ3:** How robust is the common offline popularity-calibration measurement framework under different levels of item familiarity, list composition, and available user-history?
- **RQ4:** How do computationally assigned popularity labels align with users' own popularity judgments?

By addressing these questions, this paper contributes to a more critical understanding of popularity calibration as both an offline measurement framework and a user-facing construct. We examine whether users perceive the intended differences between differently calibrated lists, whether calibrated lists are actually preferred, how sensitive calibration estimates are to incomplete user-history information, and whether computational popularity labels align with users' own judgments of item popularity.

## 2 RELATED WORK

### 2.1 Foundations and extensions of popularity calibration

Abdollahpouri et al. [2] introduced popularity calibration as a user-centered debiasing method designed to align the distribution of head, mid, and tail items in personalized recommendation lists. In contrast to global popularity-bias measures, popularity calibration evaluates whether the popularity distribution of a recommendation list matches the popularity distribution of a user's past interactions. This makes calibration a user-centered metric: a list dominated by popular items may be calibrated for one user, but miscalibrated for another.

Subsequent research has applied and extended this framework across different domains and evaluation settings. In the movie streaming sector, Klimashevskaia et al. [14] found that popularity calibration can positively affect consumer fairness, although its effectiveness for provider fairness is more limited. Within the music, movie, and anime domains, Kowald et al. [15] demonstrated that calibration and miscalibration can be used as metrics to measure popularity bias in recommender systems. Related work has also used calibration-based metrics to examine whether popularity bias affects users of different genders differently in music recommendation [20]. More recently, Forster et al. [12] combined popularity calibration with context-aware mechanisms to improve consumer-side fairness in point-of-interest recommendation.

Although popularity calibration is often evaluated through offline metrics, recent work has started to examine calibration and related bias-mitigation approaches in user-facing settings [3, 13, 28]. This line of work is important because calibration can change the composition of recommendation lists in ways that are measurable computationally, but not necessarily noticed, understood, or valued by users. For example, Alves et al. [3] found that users do not necessarily perceive calibrated recommendations as fairer unless the system explicitly explains the purpose of calibration. These findings suggest that the computational value of calibration and its perceived value for users should be treated as related but distinct questions.

A second open issue concerns the robustness of the calibration estimate itself. Popularity calibration depends not only on the recommendation list, but also on how item popularity is computed, how popularity categories are defined, and how the user's historical profile is represented. This is particularly important when calibration is interpreted as a user-centered metric, because missing, unmatched, or temporally uneven interaction histories may change the target distribution against which recommendations are evaluated.

### 2.2 Perceived vs. computational bias in recommendation

A growing body of work has examined whether computational measures of popularity bias, fairness, or miscalibration correspond to users' perceptions. This distinction is particularly relevant for music recommender systems, where users may judge recommendations not only by their computed popularity distribution, but also through artist recognition, track familiarity, and perceived fit with their music taste.

Ferwerda et al. [11] explored how users perceive popularity and fairness in music recommendations. Their findings suggest that users often do not recognize differences between presented popularity distributions, and that perceived popularity is more strongly associated with artists than with individual songs. This indicates that differences in computational popularity bias may not always be readily apparent to users.

Closely related to the present work, Lesota et al. [19] compared computational and perceived popularity miscalibration in music recommendation. They showed that JSD, a computational metric for popularity miscalibration, can correlate with user perception of recommendation lists. However, item familiarity remained an important limitation: when users are unfamiliar with recommended items, their ability to judge popularity or miscalibration is constrained. These studies motivate a more detailed examination of the conditions under which computational and perceived popularity calibration align.

This creates a methodological tension for user studies. Unfamiliar recommendations may be closer to discovery-oriented recommendation scenarios, but they make it difficult to ask users to judge popularity, liking, or fit. Familiar items, in contrast, provide a cleaner setting for studying whether users can perceive differences in popularity composition, but may also increase the role of recognition and prior liking. The present work explicitly focuses on this tension by constructing lists from recently consumed tracks.

### 2.3 User evaluation and awareness in music recommender systems

Music recommendation is a central domain for studying popularity bias and calibration because recommendation lists can affect both listener experience and exposure for artists or tracks [16, 19, 20]. Prior work has used calibration-based metrics to analyze popularity bias in music recommender systems [15, 20]. At the same time, music recommendation makes user-centered evaluation difficult: users may respond strongly to whether they recognize a track, whether they like the artist, or whether the list fits their taste [11]. Therefore, a list that is computationally calibrated is not necessarily experienced as better by the user.

Complementary insights are provided by Dinnissen et al. [7], who conducted a mixed-methods user study combining qualitative and quantitative methods, including think-aloud protocols, interviews, observed user behavior, and questionnaires. Their work highlights that users may reflect on popularity, diversity, and fairness in music recommendation, but that such reflection does not necessarily translate directly into different selection behavior. Dokoupil et al. [8, 9] further examined user perceptions of diversity and system objectives, emphasizing the nuanced ways users interpret and respond to system-driven fairness and diversity cues.

Our work builds on these foundations by studying popularity calibration under two conditions that are central to music recommendation but remain methodologically challenging: item familiarity and incomplete user history. Specifically, we examine whether users perceive differences between lists with different popularity compositions, whether calibrated lists are preferred over high- or low-popularity lists, whether offline calibration estimates remain robust under different familiarity and history-availability conditions, and whether computational popularity labels align with users' own judgments of item popularity.

## 3 METHODS AND MATERIALS

We address the research questions by conducting a user study in which participants evaluated music track lists with respect to familiarity, liking, perceived popularity, and perceived fit with their taste. We then compare these user judgments with offline measurements of popularity calibration. The study setup is designed to control two issues that complicate user studies on popularity calibration: variation in item familiarity and incomplete availability of user-history information. In this section, we describe how popularity calibration is measured, how item popularity labels are assigned, how the track lists are generated, which questions were posed, and how the resulting data are analyzed.

### 3.1 Popularity calibration measurement

A recommender is considered popularity-calibrated when, for a given user, the popularity distribution of the recommendation list matches the popularity distribution of the user's consumption history. The degree of popularity calibration is commonly evaluated at the user level as the divergence between two discrete probability distributions and then aggregated across users.

Following previous work, we distinguish between three item popularity categories, $c \in C$, where

$$C = \{\texttt{HighPop}, \texttt{MidPop}, \texttt{LowPop}\}.$$

We denote by $H_u$ the three-bin popularity distribution of the consumption history of user $u$, and by $R_u$ the corresponding three-bin popularity distribution of the recommendation list shown to user $u$. We quantify popularity miscalibration using Jensen-Shannon Divergence (JSD), shown in Eq. 1, where $H_u(c)$ is the proportion of items from category $c$ in the user's history distribution, and $R_u(c)$ is the corresponding proportion in the recommendation list. Lower JSD values indicate stronger popularity calibration, whereas higher values indicate stronger popularity miscalibration. As usual, terms with zero probability are treated as contributing zero to the sum.

$$\begin{aligned} JSD(H_u, R_u) = \frac{1}{2} \sum_c H_u(c) \cdot \log_2 \frac{2H_u(c)}{H_u(c) + R_u(c)} + \\ \frac{1}{2} \sum_c R_u(c) \cdot \log_2 \frac{2R_u(c)}{H_u(c) + R_u(c)} \end{aligned} \tag{1}$$

### 3.2 Dataset and popularity labeling

We conduct our study in the domain of music recommendation and use the LFM-2B dataset [26] to assign popularity labels to music tracks. The dataset contains user listening events, or scrobbles, from Last.fm, spanning from February 2005 to March 2020. We process the data in several steps to obtain a collection of tracks with reliable popularity estimates.

First, we match tracks to Spotify URIs using relaxed artist and track-name matching. This step helps filter out listening events involving non-music content, such as podcasts and audiobooks, which may sometimes be tracked on Last.fm. Since relaxed matching can map remixes or less popular versions of a track to the source track, we apply a stricter filtering step later. Second, we assign raw popularity values to tracks based on the number of unique users who interacted with each track. Third, following previous work [2, 17], we assign popularity labels by ranking tracks according to their popularity. The most popular tracks, which together account for 20% of the cumulative user-track popularity mass, are labeled `HighPop`. Analogously, the least popular tracks that jointly account for 20% of the total popularity mass are labeled `LowPop`, and the remaining tracks are labeled `MidPop`.

Finally, for the study itself, we only consider tracks whose artist name and title exactly match the metadata associated with the matched Spotify URI. This allows us to be more confident about

the popularity labels of the tracks shown to participants, while still allowing the broader dataset to contribute to the overall popularity distribution. After this filtering procedure, the collection contains approximately 3.4M unique tracks.

### 3.3 Naive recommender

Previous work repeatedly relies on established recommender algorithms to provide participants with recommendation lists for evaluation [11]; for example, Lesota et al. [19] generated personalized recommendations for each participant using several recommender algorithms to analyze the relation between user-perceived and computational popularity miscalibration. While this approach is representative of realistic recommendation scenarios, it can introduce limitations to user studies. In particular, two challenges are relevant to the present work: limited variability between recommendation lists and varying levels of participants' familiarity with recommended items.

Traditional recommender algorithms often produce lists biased toward more popular items [18, 20]. While the degree of bias varies between algorithms, some list types, such as lists dominated by niche items, may remain underrepresented. This can limit the range of popularity compositions available for user evaluation. A second challenge is that participants may be unfamiliar with recommended items because these items are drawn from the full catalog rather than from previously experienced tracks. This makes it difficult for participants to judge whether a track is popular, whether they like it, or whether it fits their taste. In some study designs, this issue is addressed by allowing participants to consume or preview the items, for example, by listening to audio snippets. However, this is not always desirable if the study focuses on participants' existing recognition and recollection of items, or if the researchers want to avoid introducing additional content-based effects or causing excessive participant fatigue.

In this study, we therefore use a controlled naive recommender. We use the term naive recommender to refer to a controlled list generator rather than a competitive recommendation algorithm. The goal is not to maximize recommendation accuracy, but to construct personalized lists with predefined popularity compositions while keeping item familiarity high. The naive recommender composes lists from tracks in the participant's recent listening history according to a predefined popularity distribution.

We consider three list types:

- `LowPop-Heavy`— a list of 10 random items from the participant's recent listening history containing as many `LowPop` items as possible. If the participant consumed fewer than 10 `LowPop` items, the remaining positions are filled with `MidPop` items. This list type represents the most niche part of the participant's recent listening history from the perspective of computationally assigned popularity labels.
- `HighPop-Heavy`— a list of 10 random items from the participant's recent listening history containing as many `HighPop` items as possible. If the participant consumed fewer than 10 `HighPop` items, the remaining positions are filled with `MidPop` items. This list type represents the most mainstream part of the participant's recent listening history from the perspective of computationally assigned popularity labels.
- `Calibrated`— a list of 10 items representing the popularity distribution of the participant's recent listening history. First, the popularity distribution of the participant's recent history is computed based on computationally assigned popularity labels. Then, this distribution is approximated as a discrete 10-item list. The required number of items from each popularity category is sampled from the participant's recent listening history.

Table 1. List-level evaluation questions. Responses were given on a five-point Likert scale.

| Code | Question |
|---|---|
| $Know_l$ | "The list contains tracks I know" |
| $Like_l$ | "I like this list of tracks" |
| $Pop._l$ | "The list contains a lot of popular items" |
| $Taste_l$ | "The list represents my music taste well" |
| $Unpop._l$ | "The list contains a lot of unpopular items" |

Table 2. Track-level evaluation questions and response options.

| Code | Statement | Response options |
|---|---|---|
| $Know_{tr}$ | "I know this track" | "no" / "yes" |
| $Like_{tr}$ | "I like this track" | "dislike" / "neutral" / "like" |
| $Pop._{tr}$ | "Among the music I listen to, this track is closer to" | "less popular" / "average" / "more popular" |

### 3.4 Capturing participants' perception

Each participant was presented with six personalized lists of music tracks: two `HighPop-Heavy` lists, two `LowPop-Heavy` lists, and two `Calibrated` lists. The order of the six lists was randomized for each participant. Participants evaluated each list both at the list level and at the track level. The list-level questions are shown in Table 1, and the track-level questions are shown in Table 2.

At the track level, participants first indicated whether they knew each track. The liking and perceived-popularity questions were only shown for tracks that participants indicated as familiar. This design avoids asking participants to judge the popularity or appeal of tracks they did not recognize. For the analysis, track-level responses were aggregated into counts per list. In particular, $Know_{tr}$ denotes the number of known tracks in a list, $Like_{tr}$ denotes the number of familiar tracks marked as liked, $H.Pop._{tr}$ denotes the number of familiar tracks judged as closer to more popular on the spectrum of popularity of the participant's music taste, and $L.Pop._{tr}$ denotes the number of familiar tracks judged as closer to less popular among the participant's usually listened music.

### 3.5 Participants

We recruited participants through Prolific[1]. Eligible participants were identified through a pre-screening study as active Last.fm users. In total, 80 participants were recruited for the main study. At the time of participation, participants had to satisfy the following inclusion criteria:

- being an active user of at least one music streaming platform (self-reported);
- residing in the United States;
- having recorded at least 60 listening events on Last.fm during the previous 30 days;
- during the previous seven weeks, having listened to at least 60 distinct tracks that could be matched to the dataset used for popularity labeling.

For each participant, we retrieved up to 4000 of their most recent Last.fm listening events through the Last.fm API[2], restricted to events no older than seven weeks before the start of the study. Because

[1]https://www.prolific.com/
[2]https://www.last.fm/api

the study remained open for one week, this ensured that every track shown to a participant had been listened to by that participant at least once, no more than eight weeks before participation.

Participants had an estimated completion time of 15 minutes. The observed median completion time was 14.2 minutes. Following our exclusion criterion, we removed participants who completed the study in less than 10 minutes. This resulted in a final sample of 60 participants with an average age of 31.4 ($std = 9.5$) and the gender distribution over Diverse/Female/Male as 5/27/28. The `HighPop-Heavy` lists created for the participants contain on average 9 `HighPop` items with median of 10. Out of 60 participants 11 had fewer than 10 `HighPop` items in their listening histories, of which 2 consumed no `HighPop` items at all (therefore the two received `MidPop` items in their `HighPop-Heavy` lists). All `LowPop-Heavy` lists created for the participants had all 10 items from the `LowPop` category.

### 3.6 Ethical considerations

The study did not undergo formal review by an institutional ethics board, as no applicable institutional ethics-review procedure was available for this type of study. We nevertheless followed standard ethical research practices and Prolific's requirements for participant recruitment, study description, informed participation, compensation, and withdrawal. Participants were informed about the purpose of the study, the type of data used, the expected duration of the task, and their right to discontinue participation without adverse effects. Participation was voluntary, and participants were compensated through Prolific.

Because the study relies on participants' recent Last.fm listening histories, we treated listening data as potentially identifying behavioral data. The data were used only to construct personalized study lists, assign popularity labels, and conduct the analyses reported in this paper. Participant records were pseudonymized for analysis, and results are reported only in aggregate. We do not disclose individual participants' listening histories, Last.fm profiles, or item-level responses in a form that would allow participants to be identified.

### 3.7 Analysis workflow

For each list, we compute i) track-level and ii) list-level perceptual indicators. Track-level indicators are aggregated as counts out of 10, including the number of tracks known by the participant, liked by the participant, judged as highly popular, and judged as less popular. List-level indicators are taken directly from the five-point Likert-scale questions.

Table 3 reports mean values by list type: `HighPop-Heavy` (*High*), `Calibrated` (*Cali*), `LowPop-Heavy` (*Low*). The descriptive values are averaged over lists of each type.

We use Friedman tests to test for overall differences between the three list types and Wilcoxon signed-rank tests for pairwise comparisons. We apply a Bonferroni-corrected significance threshold of $\alpha = 0.05/36$ across the tests reported in Table 3.

To analyze the relation between computational popularity miscalibration and participants' perception, we compute repeated-measures correlations between JSD and list-level perceptions of popularity and unpopularity. We conduct these analyses under different thresholds for item familiarity and available user history. Specifically, we compare settings in which participants are familiar with at least five or at least eight tracks in each list, and settings in which the user's popularity profile is estimated from different minimum amounts of matched user history.

Finally, to compare computational popularity labels with participant-judged popularity labels, we compute agreement between the computational category of each track and the participant's track-level judgment. We summarize this agreement using Cohen's $\kappa$ and Kendall's $\tau$, and additionally inspect category-level agreement for `LowPop`, `MidPop`, and `HighPop` items.

Table 3. Comparison of mean opinion scores over perceptual indicators. Values are averaged over lists of each type; Significant pairwise differences from HighPop-Heavy (*High*), Calibrated (*Cali*), and LowPop-Heavy (*Low*) lists are marked as $^h$, $^c$, and $^l$, respectively (Wilcoxon signed-rank test). Friedman test statistics ($\chi^2(2)$) for each indicator are reported in the last row, with significant differences marked by $^*$ (Bonferroni-corrected $\alpha = 0.05/36$).

| List Type | Track count (out of 10) | | | | Whole-list evaluation (Likert-5) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $Like_{tr}$ | $H.Pop._{tr}$ | $L.Pop._{tr}$ | $Know_{tr}$ | $Like_l$ | $Pop._l$ | $Unpop._l$ | $Know_l$ | $Taste_l$ |
| *High* | **7.96**$^l$ | **4.81**$^{cl}$ | 1.40$^{cl}$ | **9.12**$^{cl}$ | **4.51** | **4.15**$^{cl}$ | 2.06$^{cl}$ | **4.87**$^l$ | 3.83 |
| *Cali* | 7.48 | 2.89$^{hl}$ | 2.60$^{hl}$ | 8.57$^h$ | 4.46 | 3.32$^{hl}$ | 2.87$^{hl}$ | 4.71 | **3.87** |
| *Low* | 7.05$^h$ | 1.59$^{hc}$ | **3.80**$^{hc}$ | 8.47$^h$ | 4.26 | 2.41$^{hc}$ | **3.50**$^{hc}$ | 4.62$^h$ | 3.54 |
| Friedman $\chi^2(2)$ | 20.78$^*$ | 89.75$^*$ | 55.03$^*$ | 16.17$^*$ | 8.51 | 90.60$^*$ | 71.38$^*$ | 14.19$^*$ | 4.98 |

## 4 RESULTS

We organize the results according to the four research questions. First, we examine whether participants perceived differences between the three list types. Second, we analyze whether popularity-calibrated lists were preferred over lists dominated by high- or low-popularity tracks. Third, we investigate how the relation between computational popularity miscalibration and perceived list popularity changes under different familiarity and user-history conditions. Finally, we compare computationally assigned popularity labels with participants' own track-level popularity judgments.

*Perceived differences between list types.* Table 3 reports mean perceptual indicators for HighPop-Heavy, Calibrated, and LowPop-Heavy lists. The values are averaged over lists of each type. The results show that participants perceived clear differences in popularity composition between the three list types.

At the track level, HighPop-Heavy lists contained the highest number of tracks judged as highly popular ($H.Pop._{tr} = 4.81$), followed by Calibrated lists (2.89) and LowPop-Heavy lists (1.59). The reverse pattern appears for tracks judged as less popular: LowPop-Heavy lists received the highest number of low-popularity judgments ($L.Pop._{tr} = 3.80$), followed by Calibrated lists (2.60) and HighPop-Heavy lists (1.40). These differences are significant across all pairwise comparisons.

The same pattern is visible at the list level. Participants rated HighPop-Heavy lists as containing the most popular items ($Pop._l = 4.15$), followed by Calibrated lists (3.32) and LowPop-Heavy lists (2.41). Conversely, LowPop-Heavy lists were rated as containing the most unpopular items ($Unpop._l = 3.50$), followed by Calibrated lists (2.87) and HighPop-Heavy lists (2.06). The significant Friedman tests and pairwise Wilcoxon comparisons indicate that the list-generation procedure successfully produced lists that were perceived as different in their popularity composition.

It is also worth noting that familiarity was high across all conditions. On average, participants reported knowing more than eight tracks per list in all three conditions. However, HighPop-Heavy lists were still perceived as slightly more familiar than the other two list types, both at the track level ($Know_{tr} = 9.12$) and at the list level ($Know_l = 4.87$). This suggests that constructing lists from recent listening histories reduced unfamiliarity substantially, but did not fully eliminate familiarity differences between popularity conditions.

*Preference for calibrated lists.* Although participants perceived clear popularity differences between list types, these differences did not translate into a clear preference for calibrated lists. At the list level, liking was high for all three list types: HighPop-Heavy ($Like_l = 4.51$), Calibrated (4.46), and LowPop-Heavy (4.26). The Friedman test for list liking was not significant after correction. Similarly, perceived taste fit did not differ significantly between the list types, with Calibrated

Table 4. Repeated measures correlation with popularity miscalibration (JSD). Every participant is familiar with at least 5 tracks in each list. All three types of lists are included.

| History matched | # participants | # lists | $r_{rm}(Pop._l)$ | $p$ | $r_{rm}(Unpop._l)$ | $p$ |
|---|---|---|---|---|---|---|
| 60+ | 60 | 345 | 0.151 | 0.010 | −0.168 | 0.004 |
| 90+ | 49 | 283 | 0.185 | 0.004 | −0.184 | 0.005 |
| 120+ | 41 | 235 | 0.186 | 0.009 | −0.192 | 0.007 |

Table 5. Repeated measures correlation with popularity miscalibration (JSD). Every participant is familiar with at least 8 tracks in each list. All three types of lists are included.

| History matched | # participants | # lists | $r_{rm}(Pop._l)$ | $p$ | $r_{rm}(Unpop._l)$ | $p$ |
|---|---|---|---|---|---|---|
| 60+ | 56 | 253 | 0.218 | 0.002 | −0.194 | 0.006 |
| 90+ | 45 | 199 | 0.278 | $< .001$ | −0.189 | 0.019 |
| 120+ | 37 | 162 | 0.28 | 0.001 | −0.211 | 0.018 |

lists showing only a small numerical advantage ($Taste_l = 3.87$) over `HighPop-Heavy` (3.83) and `LowPop-Heavy` (3.54) lists.

At the track level, participants liked more tracks in `HighPop-Heavy` lists ($Like_{tr} = 7.96$) than in `LowPop-Heavy` lists (7.05), while `Calibrated` lists fell between them (7.48). This suggests that popularity-calibrated lists were not clearly preferred over popularity-miscalibrated lists. Instead, when all lists were constructed from recently consumed and therefore familiar tracks, participants generally liked the lists regardless of their calibration condition. This result qualifies the user-centered interpretation of popularity calibration: users can perceive differences in popularity composition, but this does not necessarily mean that they prefer the calibrated condition.

*Relation between JSD and perceived popularity.* Tables 4 and 5 report repeated-measures correlations between JSD and list-level perceptions of popularity and unpopularity when all three list types are included.

When participants were familiar with at least five tracks in each list, the correlations between JSD and perceived list popularity were positive but weak, ranging from $r = .151$ with at least 60 matched history items to $r = .186$ with at least 120 matched history items. The correlations between JSD and perceived list unpopularity were negative and similarly weak, ranging from $r = -.168$ to $r = -.192$.

When the familiarity threshold was increased to at least eight known tracks per list, the relation between JSD and perceived popularity became somewhat stronger. Correlations with perceived list popularity increased from $r = .218$ to $r = .280$ across the three history thresholds. Correlations with perceived list unpopularity remained negative and modest, ranging from $r = -.194$ to $r = -.211$. These results suggest that the relation between computational miscalibration and perceived popularity becomes clearer when participants are more familiar with the evaluated tracks. However, when all three list types are considered together, the overall correlations remain modest.

*Effect of list composition on JSD correlations.* A clearer pattern appears when the analysis is restricted to `HighPop-Heavy` and `Calibrated` lists.

As shown in Tables 6 and 7, correlations between JSD and perceived list popularity are substantially stronger in this setting. With a familiarity threshold of at least five known tracks per list, correlations range from $r = .526$ to $r = .550$. With a stricter threshold of at least eight known

Table 6. Repeated measures correlation with popularity miscalibration (JSD). Every participant is familiar with at least 5 tracks in each list. Only HighPop-Heavy and Calibrated lists are included.

| History matched | # participants | # lists | $r_{rm}(Pop._l)$ | $p$ | $r_{rm}(Unpop._l)$ | $p$ |
|---|---|---|---|---|---|---|
| 60+ | 60 | 235 | 0.526 | $< .001$ | $-0.495$ | $< .001$ |
| 90+ | 49 | 193 | 0.540 | $< .001$ | $-0.493$ | $< .001$ |
| 120+ | 41 | 161 | 0.550 | $< .001$ | $-0.478$ | $< .001$ |

Table 7. Repeated measures correlation with popularity miscalibration (JSD). Every participant is familiar with at least 8 tracks in each list. Only HighPop-Heavy and Calibrated lists are included.

| History matched | # participants | # lists | $r_{rm}(Pop._l)$ | $p$ | $r_{rm}(Unpop._l)$ | $p$ |
|---|---|---|---|---|---|---|
| 60+ | 52 | 173 | 0.511 | $< .001$ | $-0.469$ | $< .001$ |
| 90+ | 41 | 135 | 0.537 | $< .001$ | $-0.438$ | $< .001$ |
| 120+ | 33 | 111 | 0.523 | $< .001$ | $-0.402$ | $< .001$ |

tracks, correlations remain similarly strong, ranging from $r = .511$ to $r = .537$. The corresponding correlations with perceived list unpopularity are negative and also substantial.

This indicates that JSD aligns more clearly with user perception when the comparison is between mainstream-heavy and calibrated lists. When LowPop-Heavy lists are included, the relation becomes weaker. One likely reason is that JSD measures the magnitude of divergence between the list and the user's historical popularity profile, but does not by itself encode whether the divergence comes from a shift toward more popular or less popular items. Thus, JSD can indicate that a list is miscalibrated, but additional information is needed to interpret the direction of the miscalibration.

*Influence of available user history.* The history-threshold comparisons show that the relation between JSD and perceived popularity is sensitive to how much matched user history is available. In the all-list setting, increasing the minimum amount of matched history from 60 to 90 or 120 tracks generally increases the correlation between JSD and perceived popularity, although the effect remains modest. This suggests that popularity calibration estimates become somewhat more aligned with user perception when the user's target popularity profile is based on more matched listening history.

However, the pattern is not uniformly stronger across all settings. In the HighPop-Heavy–Calibrated comparison, correlations are already strong at the 60-track threshold and change only slightly as the amount of matched history increases. This suggests that the effect of available user history depends on list composition. Additional matched history may matter most when the analysis includes multiple forms of miscalibration, whereas the distinction between high-popularity-heavy and calibrated lists is already captured relatively clearly with less history.

*Alignment between computational and participant-judged popularity labels.* Finally, we compare computational popularity labels with participants' track-level popularity judgments. Agreement between the two categorizations was weak overall, with a mean Cohen's $\kappa$ of .21 and a mean Kendall's $\tau$ of .36. Agreement was higher for the extreme categories than for the middle category: 62% of participant-assigned low-popularity judgments matched the corresponding computational label, and 58% of high-popularity judgments matched the computational label. In contrast, only 24% of participant-assigned mid-popularity judgments matched the computational mid-popularity category.

This weak alignment indicates that computational popularity labels and participant-judged popularity labels should not be treated as interchangeable. Participants appear to identify relatively popular and relatively unpopular tracks more consistently than mid-popularity tracks, but their judgments do not fully correspond to the popularity bins used for offline calibration. This complicates the interpretation of popularity calibration as a user-centered metric, because the target distribution is based on computational categories that only partially align with users' own popularity perceptions.

*Summary.* Overall, the results show that participants perceive the intended differences between `HighPop-Heavy`, `Calibrated`, and `LowPop-Heavy` lists. However, calibrated lists are not clearly preferred over high- or low-popularity-heavy lists. The relation between JSD and perceived popularity is present but sensitive to item familiarity, list composition, and the amount of matched user history. Finally, the weak agreement between computational and user-judged popularity labels raises questions about how directly offline popularity categories can represent users' own perceptions of popularity.

## 5 DISCUSSION

Our results provide a nuanced view of popularity calibration as a user-centered metric in music recommendation. On the one hand, participants clearly perceived the intended differences between lists with different popularity compositions. Lists constructed to be high-popularity-heavy were judged as more popular, while low-popularity-heavy lists were judged as more unpopular. This suggests that users can perceive differences in popularity composition when the presented tracks are sufficiently familiar. On the other hand, these perceived differences did not translate into a clear preference for popularity-calibrated lists. This distinction is important: a calibration metric may capture a meaningful computational property of a recommendation list, but this does not automatically mean that users experience calibrated lists as more useful, more enjoyable, or more representative of their taste.

### 5.1 Popularity calibration as a user-centered metric

Popularity calibration is often motivated as a user-centered alternative to global popularity-bias measures [2, 5]. Rather than asking whether a system recommends too many popular items overall, calibration asks whether the popularity profile of a recommendation list matches the user's historical consumption profile. This makes the metric attractive because it recognizes that popularity preference may differ between users. A mainstream-oriented listener may be expected to receive more popular tracks, while a listener with a stronger long-tail profile may reasonably receive more niche tracks.

However, our results show that this user-centered framing should be interpreted carefully. Participants perceived differences between high-popularity-heavy, low-popularity-heavy, and calibrated lists, but they did not clearly prefer the calibrated lists. List-level liking was high across all three conditions, and perceived taste fit did not significantly differ between list types. This suggests that calibration may describe a certain kind of alignment with past behavior, but not necessarily the kind of alignment users value most when evaluating music recommendations.

One possible explanation is that our study deliberately controlled item familiarity by constructing lists from recently consumed tracks. This design makes it easier to study whether users perceive differences in popularity composition, but it also means that all lists contain relatively familiar and recently relevant items. Under these conditions, even a popularity-miscalibrated list may still

be liked because it consists of tracks the user knows and has listened to recently. This interpretation aligns with prior work suggesting that perceived popularity and user evaluations of music recommendations can be strongly affected by familiarity and recognition [7, 11].

### 5.2 What should be the target of calibration?

A central question raised by our findings is what should count as the target distribution for popularity calibration. In standard offline evaluation, the target is usually derived from the user's historical interactions [2, 5]. However, user histories are not neutral representations of preference. They depend on the time span covered by the dataset, the availability of item metadata, the platform from which interactions are collected, and the filtering decisions made during preprocessing.

This is particularly important in music recommendation. A user's complete listening history may span many years, but recent listening may better represent current taste. Conversely, a short recent history may be too narrow to capture stable preference patterns. Prior work on music preferences and mainstreaminess already shows that music consumption can vary across users, countries, and temporal contexts [4, 21, 22]. The issue becomes even more complex when parts of the listening history cannot be matched to the item catalog used for popularity estimation. In such cases, the target popularity profile is not simply "the user's history," but the subset of the user's history that remains observable after matching, filtering, and popularity labeling.

Our correlation analyses show that the relation between JSD and perceived list popularity changes with the amount of matched user history. This indicates that popularity calibration estimates are sensitive to how the user's historical profile is represented. As a result, calibration should not be treated as a purely objective property of a recommendation list. It is also a property of the data representation used to construct the user's target profile.

This has implications for offline evaluation. If different datasets cover different temporal spans, or if different users have histories of very different lengths, then the meaning of calibration may vary across users and domains. A "complete history" may represent years of listening for one user and weeks of activity for another. Therefore, future work should be more explicit about what portion of user history is treated as representative of popularity preference, and why.

### 5.3 What does "popular" mean to users?

Our results also show weak agreement between computational popularity labels and participant-judged popularity labels. This raises a second construct-validity issue: offline metrics usually define popularity through aggregate interaction data, whereas users may judge popularity through different cues. For example, a participant may judge a track as popular because they recognize the artist, because the song is widely known in their social context, or because it belongs to a mainstream genre. Conversely, a track may have many interactions in the dataset but still not be perceived as popular by a specific listener.

This concern is consistent with prior work showing that perceived popularity in music recommendation may be associated more strongly with artists than with individual songs [11]. It also connects to broader work on music mainstreaminess, where popularity can be defined at different levels, such as artist playcounts, listener counts, global mainstreaminess, or country-specific mainstreaminess [4]. In our study, computational labels and user judgments aligned only weakly overall, with stronger agreement for the extreme categories than for the middle category. This suggests that users may be relatively able to identify very popular or very niche tracks, while mid-popularity items are harder to classify consistently.

For popularity calibration, this matters because the metric depends on computational popularity bins. If these bins only partially align with users' own judgments, then a list that is computationally calibrated may not be perceived as calibrated by the user. This does not invalidate computational

calibration, but it clarifies what it measures: alignment with a dataset-derived popularity distribution, not necessarily alignment with the user's subjective understanding of popularity.

### 5.4 Direction of miscalibration

Another implication concerns the direction of miscalibration. JSD measures the divergence between the user's historical popularity distribution and the recommendation-list distribution [24]. It captures the magnitude of miscalibration, but not whether the list shifts the user toward more popular or less popular items. Our results show that JSD correlates more strongly with perceived popularity when the comparison is between high-popularity-heavy and calibrated lists. When low-popularity-heavy lists are included, the correlations become weaker.

This suggests that the interpretation of JSD depends on list composition. A high JSD value can indicate that a list is miscalibrated, but it does not explain how it is miscalibrated. For user-centered evaluation, this distinction is important. A list that is too mainstream and a list that is too niche may have similar JSD values, but they may be perceived very differently by users and may have different implications for personalization, discovery, and fairness.

Future work should therefore consider complementing symmetric divergence-based calibration metrics with directional indicators. For example, reporting whether a list shifts the user toward more high-popularity or low-popularity items would make calibration results easier to interpret. This would also help distinguish between calibration as a personalization objective and calibration as a fairness or exposure-related objective.

### 5.5 Calibration, fairness, and user value

Popularity calibration is often connected to fairness because reducing popularity bias may improve exposure for less popular items or providers [1, 2, 20]. However, our results underline that fairness value and user-perceived value are not the same. A calibrated list may be desirable from the perspective of reducing popularity distortion, but users may not necessarily prefer it or perceive it as a better fit.

This distinction is consistent with recent work showing that users do not always perceive calibrated or fairness-oriented recommendation interventions in the way system designers intend [3, 13, 28]. It also connects to broader work on fairness in recommender systems, which emphasizes that fairness is a multi-stakeholder and context-dependent concept [6, 10, 27].

This means that the purpose of calibration should be made explicit. If the goal is personalization, then calibration should be evaluated in terms of whether it improves perceived taste fit, satisfaction, or long-term usefulness. If the goal is fairness or exposure diversity, then calibration may be valuable even when users do not immediately prefer calibrated lists. These are different normative goals, and the same metric should not be assumed to satisfy all of them equally.

### 5.6 Limitations

Several limitations should be considered when interpreting our findings. First, our naive recommender was designed to control list composition and item familiarity rather than to compete with state-of-the-art recommender algorithms. This design is useful for isolating the role of popularity composition, but it does not fully reflect how users encounter new recommendations in real systems.

Second, because lists were constructed from recently consumed tracks, familiarity was high across all conditions. This was intentional, but it may also reduce the observable value of calibration. Users may like all lists because the items are familiar and recently relevant, regardless of whether the list is calibrated. Future studies could systematically vary familiarity to examine when calibration becomes more or less valuable.

Third, our computational popularity labels are derived from the available dataset and preprocessing pipeline. Different datasets, time windows, or popularity definitions could produce different labels and, therefore, different calibration estimates. This is not only a limitation of the present study but also an important methodological issue for popularity calibration research more broadly.

Finally, our study focuses on perceived popularity, liking, familiarity, and taste fit. Other outcomes, such as perceived fairness, discovery value, novelty, long-term satisfaction, or artist-level exposure, may reveal additional aspects of the value of calibration. Prior work on diversity, nudging, and user perceptions of recommendation objectives suggests that such outcomes may be important for understanding the broader user experience of music recommender systems [8, 9, 23, 25].

### 5.7 Implications

The main implication of our study is that popularity calibration should not be treated as self-evidently user-centered simply because it uses the user's historical profile as a reference point. The metric depends on how that profile is constructed, how popularity is defined, and whether computational popularity categories correspond to users' own judgments. Our findings suggest that users can perceive differences in popularity composition, but that calibrated lists are not necessarily preferred, and that calibration estimates are sensitive to familiarity, list composition, and available user history.

For researchers, this means that popularity calibration should be reported with greater attention to measurement choices: the time span of user history, the amount of matched history, the popularity-binning strategy, and whether miscalibration is driven by shifts toward more popular or less popular items. For system designers, the findings suggest that calibration may be useful as one component of recommendation evaluation, but should be interpreted alongside user-facing measures such as liking, familiarity, perceived taste fit, and perceived discovery value.

## 6 CONCLUSION

Popularity calibration is commonly treated as a user-centered way to evaluate and mitigate popularity bias in recommender systems. In this paper, we examined this assumption in the context of music recommendation, focusing on item familiarity, list composition, and incomplete user-history information. We conducted a user study in which participants evaluated track lists with different popularity compositions, generated from their recent listening histories, and compared their judgments with offline popularity-calibration estimates.

With respect to **RQ1**, our results show that participants perceived clear differences between lists generated with different calibration targets. High-popularity-heavy lists were perceived as more popular, low-popularity-heavy lists were perceived as more unpopular, and calibrated lists generally fell between these two conditions. This indicates that users can perceive differences in popularity composition when the recommended items are sufficiently familiar.

For **RQ2**, however, these perceptual differences did not translate into a clear preference for calibrated lists. List liking and perceived taste did not significantly differ between the list types. This suggests that popularity calibration captures a meaningful property of recommendation lists, but that this property should not be equated directly with user appreciation or perceived preference.

Regarding **RQ3**, we found that the relation between JSD-based popularity miscalibration and perceived list popularity depends on familiarity, list composition, and the amount of available matched user history. JSD aligned more strongly with perceived popularity when comparing high-popularity-heavy and calibrated lists, but the relation became weaker when low-popularity-heavy lists were included. This highlights that offline calibration estimates are sensitive to both the available user-history information and the direction of miscalibration.

Finally, for **RQ4**, the agreement between computational popularity labels and participant-judged popularity labels was weak ($\kappa = .21$, $\tau = .36$). This suggests that dataset-derived popularity categories only partially capture users' perception of item popularity. This has implications for the use of popularity calibration as a user-centered metric, because the metric depends on computational labels that may not fully align with users' subjective popularity judgments.

Overall, our findings call for a more careful interpretation of popularity calibration in music recommender systems. Calibration remains useful as an offline measurement framework, but its user-centered value depends on how popularity is defined, how user history is represented, and whether calibrated lists are actually perceived and valued by users. Future work should further investigate directional calibration measures, alternative representations of user-history targets, and user-facing outcomes beyond liking, such as discovery value, perceived fairness, and long-term satisfaction.

## REFERENCES


[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura (Eds.). ACM, 726–731. https://doi.org/10.1145/3383313.3418487

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. 2021. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 119–129. https://doi.org/10.1145/3450613.3456821

[3] Gabrielle Alves, Dietmar Jannach, Rodrigo Ferrari de Souza, and Marcelo Garcia Manzato. 2024. User Perception of Fairness-Calibrated Recommendations. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2024, Cagliari, Italy, July 1-4, 2024*. ACM, 78–88. https://doi.org/10.1145/3627043.3659558

[4] Christine Bauer and Markus Schedl. 2019. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE* 14, 6 (2019), 1–36. https://doi.org/10.1371/journal.pone.0217389

[5] Diego Corrêa da Silva and Dietmar Jannach. 2026. Calibrated Recommendations: Survey and Future Directions. *Trans. Recomm. Syst.* 4, 3 (2026), 43:1–43:32. https://doi.org/10.1145/3789266

[6] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogín, Alessandro Difonzo, and Dario Zanzonelli. 2024. Fairness in recommender systems: research landscape and future directions. *User Model. User Adapt. Interact.* 34, 1 (2024), 59–108. https://doi.org/10.1007/S11257-023-09364-Z

[7] Karlijn Dinnissen, Shah Noor Khan, Hanna Hauptmann, Eelco Herder, and Judith Masthoff. 2025. The Role of Fairness and Diversity in User Choices and Perceptions of Music Playlists. In *Proceedings of the 36th ACM Conference on Hypertext and Social Media*. 48–57.

[8] Patrik Dokoupil, Ludovico Boratto, and Ladislav Peska. 2024. User Perceptions of Diversity in Recommender Systems. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 212–222. https://doi.org/10.1145/3627043.3659555

[9] Patrik Dokoupil, Ludovico Boratto, and Ladislav Peska. 2025. How Do Users Perceive Recommender Systems' Objectives?. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys 2025, Prague, Czech Republic, September 22-26, 2025*, Mária Bieliková, Pavel Kordík, Markus Schedl, Marco de Gemmis, Sole Pera, Rodrigo Alves, Olivier Jeunen, and Vito Ostuni (Eds.). ACM, 165–176. https://doi.org/10.1145/3705328.3748066

[10] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Found. Trends Inf. Retr.* 16, 1-2 (2022), 1–177. https://doi.org/10.1561/1500000079

[11] Bruce Ferwerda, Eveline Ingesson, Michaela Berndl, and Markus Schedl. 2023. I Don't Care How Popular You Are! Investigating Popularity Bias in Music Recommendations from a User's Perspective. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19-23, 2023*, Jacek Gwizdka and Soo Young Rieh (Eds.). ACM, 357–361. https://doi.org/10.1145/3576840.3578287

[12] Andrea Forster, Simone Kopeinik, Denis Helic, Stefan Thalmann, and Dominik Kowald. 2025. Exploring the effect of context-awareness and popularity calibration on popularity bias in poi recommendations. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 593–598.

[13] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Astrid Tessem, and Christoph Trattner. 2023. Evaluating The Effects of Calibrated Popularity Bias Mitigation: A Field Study. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song (Eds.). ACM, 1084–1089. https://doi.org/10.1145/3604915.3610637

[14] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Christoph Trattner, and Lars Skjærven. 2022. Mitigating Popularity Bias in Recommendation: Potential and Limits of Calibration Approaches. In *Advances in Bias and Fairness in Information Retrieval - Third International Workshop, BIAS 2022, Stavanger, Norway, April 10, 2022, Revised Selected Papers (Communications in Computer and Information Science, Vol. 1610)*, Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo (Eds.). Springer, 82–90. https://doi.org/10.1007/978-3-031-09316-6_8

[15] Dominik Kowald, Gregor Mayr, Markus Schedl, and Elisabeth Lex. 2023. A study on accuracy, miscalibration, and popularity bias in recommendations. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–16.

[16] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12036)*. Springer, 35–42. https://doi.org/10.1007/978-3-030-45442-5_5

[17] Oleg Lesota, Adrian Bajko, Max Walder, Matthias Wenzel, Antonela Tommasel, and Markus Schedl. 2025. Fine-tuning for Inference-efficient Calibrated Recommendations. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys 2025, Prague, Czech Republic, September 22-26, 2025*, Mária Bieliková, Pavel Kordík, Markus Schedl, Marco de Gemmis, Sole Pera, Rodrigo Alves, Olivier Jeunen, and Vito Ostuni (Eds.). ACM, 1187–1192. https://doi.org/10.1145/3705328.3759319

[18] Oleg Lesota, Stefan Brandl, Matthias Wenzel, Alessandro B. Melchiorre, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2022. Exploring Cross-group Discrepancies in Calibrated Popularity for Accuracy/Fairness Trade-off Optimization. In *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA, 18th-23rd September 2022 (CEUR Workshop Proceedings, Vol. 3268)*, Himan Abdollahpouri, Shaghayegh Sahebi, Mehdi Elahi, Masoud Mansoury, Babak Loni, Zahra Nazari, and Maria Dimakopoulou (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3268/paper5.pdf

[19] Oleg Lesota, Gustavo Escobedo, Yashar Deldjoo, Bruce Ferwerda, Simone Kopeinik, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2023. Computational Versus Perceived Popularity Miscalibration in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). ACM, 1889–1893. https://doi.org/10.1145/3539618.3591964

[20] Oleg Lesota, Alessandro B. Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 601–606. https://doi.org/10.1145/3460231.3478843

[21] Oleg Lesota, Emilia Parada-Cabaleiro, Stefan Brandl, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2022. Traces of Globalization in Online Music Consumption Patterns and Results of Recommendation Algorithms. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, Preeti Rao, Hema A. Murthy, Ajay Srinivasamurthy, Rachel M. Bittner, Rafael Caro Repetto, Masataka Goto, Xavier Serra, and Marius Miron (Eds.). 291–297. https://archives.ismir.net/ismir2022/paper/000034.pdf

[22] Elisabeth Lex, Dominik Kowald, and Markus Schedl. 2020. Modeling Popularity and Temporal Drift of Music Genre Preferences. *Trans. Int. Soc. Music. Inf. Retr.* 3, 1 (2020), 17–30. https://doi.org/10.5334/TISMIR.39

[23] Yu Liang and Martijn C Willemsen. 2022. Exploring the longitudinal effects of nudging on users' music genre exploration behavior and listening preferences. In *Proceedings of the 16th ACM conference on recommender systems*. 3–13.

[24] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* 37, 1 (1991), 145–151. https://doi.org/10.1109/18.61115

[25] Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2024. Assessing the impact of music recommendation diversity on listeners: A longitudinal study. *ACM Transactions on Recommender Systems* 2, 1 (2024), 1–47.

[26] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*, David Elsweiler (Ed.). ACM, 337–341. https://doi.org/10.1145/3498366.3505791

[27] Markus Schedl, Navid Rekabsaz, Elisabeth Lex, Tessa Grosz, and Elisabeth Greif. 2022. Multiperspective and Multidisciplinary Treatment of Fairness in Recommender Systems Research. In *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022, Adjunct Proceedings*. ACM, 90–94. https://doi.org/10.1145/3511047.3536400

[28] Robin Ungruh, Karlijn Dinnissen, Anja Volk, Maria Soledad Pera, and Hanna Hauptmann. 2024. Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, Tommaso Di Noia, Pasquale Lops, Thorsten Joachims, Katrien Verbert, Pablo Castells, Zhenhua Dong, and Ben London (Eds.). ACM, 169–178. https://doi.org/10.1145/3640457.3688102